Co-scheduling Amdahl applications on cache-partitioned systems

Guillaume Aupy, Anne Benoit, Sicheng Dai, Loïc Pottier, Padma Raghavan, Yves Robert, Manu Shantharam

Co-scheduling Amdahl applications on cache-partitioned systems

Guillaume Aupy[a], Anne Benoit[b], Sicheng Dai[c], Loïc Pottier[b], Padma Raghavan[d], Yves Robert[b,e], Manu Shantharam[f]

[a] *Inria, Université de Bordeaux, France*
[b] *Laboratoire LIP, École Normale Supérieure de Lyon, France*
[c] *East China Normal University, China*
[d] *Vanderbilt University, Nashville TN, USA*
[e] *University of Tennessee Knoxville, USA*
[f] *San Diego Supercomputer Center, San Diego CA, USA*

Abstract

Cache-partitioned architectures allow subsections of the shared last-level cache (LLC) to be exclusively reserved for some applications. This technique dramatically limits interactions between applications that are concurrently executing on a multi-core machine. Consider n applications that execute concurrently, with the objective to minimize the makespan, defined as the maximum completion time of the n applications. Key scheduling questions are: (i) which proportion of cache and (ii) how many processors should be given to each application? In this paper, we provide answers to (i) and (ii) for Amdahl applications. Even though the problem is shown to be NP-complete, we give key elements to determine the subset of applications that should share the LLC (while remaining ones only use their smaller private cache). Building upon these results, we design efficient heuristics for Amdahl applications. Extensive simulations demonstrate the usefulness of co-scheduling when our efficient cache partitioning strategies are deployed.

Keywords: Co-scheduling; cache partitioning; complexity results.

1. Introduction

At scale, the I/O movements of High Performance Computing (HPC) applications are expected to be one of the most critical problems [Adv14]. Observations on the Intrepid machine at Argonne National Laboratory (ANL) show that I/O transfers can be slowed down up to 70% due to congestion [GAB+15]. When ANL upgraded its house supercomputer from Intrepid (Peak perf: 0.56 PFlops; peak I/O throughput: 88 GB/s) to Mira (Peak perf: 10 PFlops; peak I/O throughput: 240 GB/s), the net result for an application whose I/O throughput scales linearly (or worse) with performance was a downgrade from 160 GB/PFlop to 24 GB/PFlop!

To cope with such an imbalance (which is not expected to reduce on future platforms), a possible approach is to develop *in situ* co-scheduling analysis and data preprocessing on dedicated nodes [Adv14]. This scheme applies to data-intensive periodic workflows where data is generated by the main simulation, and parallel processes are run to process this data with the constraints that output results should be sent to disk storage before newly generated data arrives for processing. These solutions are starting to be implemented for HPC applications. Sewell et al. [SHF+15] explain that in the case of the HACC application (a cosmological code), petabytes of data are created to be analyzed later. The analysis is done by multiple independent processes. The idea of their work is to minimize the amount of data copied to I/O filesystem, by performing the analysis at the same time as HACC is running (what they call *in situ*). The main constraint is that these processes are data-intensive and are handled by a dedicated machine. Also, the execution of these processes should be done efficiently enough so that they finish before the next batch of data arrives, hence resulting in a pipelined approach. All these frameworks motivate the design of efficient co-scheduling strategies.

Email addresses: guillaume.aupy@inria.fr (Guillaume Aupy), Anne.Benoit@ens-lyon.fr (Anne Benoit), 51151500012@ecnu.cn (Sicheng Dai), Loic.Pottier@ens-lyon.fr (Loïc Pottier), padma.raghavan@vanderbilt.edu (Padma Raghavan), Yves.Robert@inria.fr (Yves Robert), shantharam.manu@gmail.com (Manu Shantharam)

One main issue of co-scheduling is to evaluate co-run degradations due to cache sharing [ZBF10]. Many studies have shown that interferences on the shared last-level cache (LLC) can be detrimental to co-scheduled applications [LK14]. Previous solutions consisted in preventing co-schedule of possibly interfering workloads, or terminating low importance applications [ZLMT14]. Lo et al. [LCG+16] recently showed experimentally that important gains could be reached by co-scheduling applications with strict cache partitioning enabled. Cache partitioning, the technique at the core of this work, consists in reserving exclusivity of subsections of the LLC of a chip multi-processor (CMP), to some of the applications running on this CMP. This functionality was recently introduced by Intel under the name *Cache Allocation Technology* [Int14]. With the advent of large shared memory multi-core machines (e.g., Sunway TaihuLight, the current #1 supercomputer uses 256-cores processor chips with a shared memory of 32GB [Don16]), the design of algorithms that co-schedule applications efficiently and decide how to partition the shared memory (seen as the cache here), is becoming critical.

In this work, we study the following problem. We are given a set of Amdahl applications, i.e., parallel applications application obeying Amdahl's speedup law [Amd67] (see Equation 1 for details). Amdahl's law has had a profound impact on the evolution of HPC [Hea15] and many scientific applications, including most Nas Parallel Benchmarks, obey this law [CE00]. We are also given a multi-core processor with a shared last-level cache LLC. How can we best partition the LLC to minimize the total execution time (or *makespan*), i.e., the moment when the last application finishes its computation. For each application, we assume that we know the number of compute operations to perform, and the miss rate on a fixed size cache. For the multi-core processor, we know its LLC size, the cost for a cache miss, the cost for a cache hit, the size of the cache and total number of processors. For the theoretical study, we assume that these processors can be shared by two applications through multi-threading [KSS12], hence we can assign a rational number of processors to each application, and this allows us to study the intrinsic complexity of co-scheduling with cache partitioning. Equipped with all these applications and platform parameters, recent work [HSPE08, RKB+09, KSS12] shows how to model the impact of cache misses and to accurately predict the execution time of an application. In this context, we make the following main contributions:

- With rational numbers of processors, we show that the co-scheduling problem is NP-complete, even when applications are perfectly parallel, i.e., their speed-up scales up linearly with the number of processors.
- With rational numbers of processors, we show several results that characterize optimal solutions, and in particular that the co-scheduling cache-partitioning problem reduces to deciding which subset of applications will share the LLC; when this subset is known, we show how to determine the optimal cache fractions and rational number of processors for perfectly-parallel applications. Furthermore, we show that all applications should finish at the same time, even if they are not perfectly parallel.
- These theoretical results guide the design of heuristics for Amdahl applications. We show through extensive simulations (using both rational and integer numbers of processors) that our heuristics greatly improve the performance of cache-partitioning algorithms, even for parallel applications obeying Amdahl's law with a large sequential fraction, hence with a limited speedup profile.

The rest of the paper is organized as follows. Section 2 provides an overview of related work. Section 3 is devoted to formally defining the framework and all model parameters. Section 4 gives our main theoretical contributions. The heuristics are defined in Section 5, and evaluated through simulations in Section 6. Finally, Section 7 outlines our main findings and discusses directions for future work.

2. Related work

Since the advent of systems with tens of cores, co-scheduling has received considerable attention. Due to lack of space, we refer to [MSM+11, DJF+15, LCG+16] for a survey of many approaches to co-scheduling. The main idea is to execute several applications concurrently rather than in sequence, with the objective to increase platform throughput. Indeed, some individual applications may well not need all available cores, or some others could use all resources, but at the price of a dramatic performance loss. In particular, the latter case is encountered whenever application speedup becomes too low beyond a given processor count.

The main difficulty of co-scheduling is to decide which applications to execute concurrently, and how many cores to assign to each of them. Indeed, when executing simultaneously, any two applications will compete for shared resources, which will create interferences and decrease their throughput. Modeling application interference is a challenging task. Dynamic schedulers are used when application behavior is unknown [QP06, TJS09]. Static schedulers aim at optimizing the sharing of the resources by relying on application knowledge such as estimated workload, speed-up profile, cache behavior, etc. One widely-used approach is to build an interference graph whose vertices are applications and whose edges represent degradation factors [JSCT08, ZHG+15, HZJ16]. This approach is interesting but hard to implement. Indeed, the interaction of two applications depends on many factors, such as their size, their core count, the memory bandwidth, etc. Obtaining the speedup profile of a single application already is difficult and requires intensive benchmarking campaigns. Obtaining the degradation profile of two applications is even more difficult and can be achieved only for regular applications. To further darken the picture, the interference graph subsumes only pairwise interactions, while a global picture of the processor and cache requirements for all applications is needed by the scheduler.

Shared resources include cache, memory, I/O channels and network links, but among potential degradation factors, cache accesses are prominent. When several applications share the cache, they are granted a fraction of cache lines as opposed to the whole cache, and their cache miss ratio increases accordingly. Multiple cache partitioning strategies have been proposed [BCSM08, GSYY09, BZF10, DFB+12]. In this paper, we focus on a static allocation of LLC cache fractions, and processor numbers, to concurrent applications as a function of several parameters (cache-miss ratio, access frequency, operation count). To the best of our knowledge, this work is the first analytical model and complexity study for this challenging problem.

3. Model

This section details platform and application parameters, and formally states the optimization problem.

Architecture. We consider a parallel platform of p homogeneous computing elements, or *processors*, that share two storage locations:

- A small storage $\mathcal{S}_s$ with low latency, governed by a LRU replacement policy, also called *cache*;
- A large storage $\mathcal{S}_l$ with high latency, also called *memory*.

More specifically, C_s (resp. C_l) denotes the size of $\mathcal{S}_s$ (resp. $\mathcal{S}_l$), and l_s (resp. l_l) the latency of $\mathcal{S}_s$ (resp. $\mathcal{S}_l$). In this work, we assume that $C_l = +\infty$. We have the relation $l_s \ll l_l$.

In this work, we consider the cache partitioning technique [Int14], where one can allocate a portion of the cache to applications so that they can execute without interference from other applications.

Applications. There are n independent parallel applications to be scheduled on the parallel platform, whose speedup profiles obey Amdahl's law [Amd67]. For an application T_i, we define several parameters:

- w_i, the number of computing operations needed for T_i;
- s_i, the sequential fraction of T_i;
- f_i, the frequency of data accesses of T_i: f_i is the number of data accesses per computing operation;
- a_i, the memory footprint of T_i.

We use these parameters to model the execution of each application as follows.

Parallel execution time. Let $Fl_i(p_i)$ be the number of operations performed by each processor for application T_i, when executed on p_i processors. According to Amdahl's speedup profile [Amd67], we have

$$Fl_i(p_i) = s_i w_i + (1 - s_i)\frac{w_i}{p_i} \tag{1}$$

The power law of cache misses. In chip multi-processors, many authors have observed that the Power Law accurately models how the cache size affects the miss rate [HSPE08, RKB+09, KSS12]. Mathematically, the power law states that if m_0 is the miss rate of a workload for a baseline cache size C_0, the miss rate m for a new cache size C can be expressed as $m = m_0 \left(\frac{C_0}{C}\right)^{\alpha}$ where α is the sensitivity factor from the Power Law of Cache Misses [HSPE08, RKB+09, KSS12] and typically ranges between 0.3 and 0.7 with an average at 0.5.

Note that, by definition, a rate cannot be higher than 1, hence we extend this definition as:

$$m = \min\left(1, m_0\left(\frac{C_0}{C}\right)^{\alpha}\right). \tag{2}$$

This formula can be read as follows: if the cache size allocated is too small, then the execution goes as if no cache was allocated, and all accesses will be misses.

Computations and data movement. We use the cost model introduced by Krishna et al. [KSS12] to evaluate the execution cost of an application as a function of the cache fraction that it has been allocated. Specifically, for each application, we define m_0, the miss rate of application T_i with a cache of size C_0 (we can also use the miss rate of applications with a cache of another fixed size). We express the execution time of T_i as a function of p_i, the number of processors allocated to T_i, and x_i, the fraction of $\mathcal{S}_s$ allocated to T_i (recall both are rational numbers). Let $Fl_i(p_i)$ be the number of operations performed by each processor for application T_i, given that the application is executed on p_i processors. We have $Fl_i(p_i) = s_i w_i + (1 - s_i)\frac{w_i}{p_i}$ according to Amdahl's speedup profile. Finally,

$$\mathcal{E}\text{xe}_i(p_i, x_i) = \begin{cases} Fl_i(p_i)\left(1 + f_i\left(l_s + l_l\right)\right) & \text{if } x_i = 0; \\ Fl_i(p_i)\left(1 + f_i\left(l_s + l_l \cdot \min\left(1, \frac{m_0}{\left(\frac{x_i C_s}{C_0}\right)^{\alpha}}\right)\right)\right) & \text{if } x_i C_s \leq a_i; \\ Fl_i(p_i)\left(1 + f_i\left(l_s + l_l \cdot \min\left(1, \frac{m_0}{\left(\frac{a_i}{C_0}\right)^{\alpha}}\right)\right)\right) & \text{otherwise.} \end{cases} \tag{3}$$

Indeed, for each operation, we pay the cost of the computing operation, plus the cost of data accesses, and by definition we have f_i accesses per operation. At each access, we pay a latency l_s, and an additional latency l_l in case of cache miss (see Equation (2)). The last case states that we cannot use a portion of cache greater than the memory footprint a_i of application T_i. This model is somewhat pessimistic: cache accesses to the same variable by two different processors are counted twice. We show in Section 6 that despite this conservative assumption (no sharing), co-scheduling can outperform classical approaches that sequentially deploy each application on the whole set of available resources.

Equation (3) calls for a few observations. For notational convenience, let $d_i = m_0\left(\frac{C_0}{C_s}\right)^{\alpha}$:

- It is useless to give a fraction of cache larger than $\frac{a_i}{C_s}$ to application T_i;
- Because of the minimum $\min\left(1, \frac{d_i}{(x_i)^{\alpha}}\right)$, either $x_i > d_i^{\frac{1}{\alpha}}$, or $x_i = 0$: indeed, if we give application T_i a fraction of cache smaller than $d_i^{\frac{1}{\alpha}}$, the minimum is equal to 1, and this fraction is wasted.

Hence, we have for all i:

$$x_i = 0 \quad \text{or} \quad d_i^{\frac{1}{\alpha}} < x_i \leq \frac{a_i}{C_s}. \tag{4}$$

Of course, if $d_i^{\frac{1}{\alpha}} \geq \frac{a_i}{C_s}$ for some application T_i, then $x_i = 0$.

We denote by $\mathcal{E}\text{xe}_i^{\text{seq}}(x_i) = \mathcal{E}\text{xe}_i(1, x_i)$ the sequential execution time of application T_i with a fraction of cache x_i.

Scheduling problem. Given n applications $T_1, \ldots, T_n$, we aim at partitioning the shared cache and assign processors so that the concurrent execution of these applications takes minimal time. In other words, we aim at minimizing the execution time of the longest application, when all applications start their execution at the same time. Formally:

Definition 1 (CoSchedCache)**.** *Given n applications $T_1, \ldots, T_n$ and a platform with p identical processors sharing a cache of size C_s, find a schedule $\{(p_1, x_1), \ldots, (p_n, x_n)\}$ with $\sum_{i=1}^{n} p_i \leq p$, and $\sum_{i=1}^{n} x_i \leq 1$, that minimizes*

$$\max_{1 \leq i \leq n} \mathcal{E}xe_i(p_i, x_i).$$

We pay particular attention in the following to *perfectly parallel* applications, i.e., applications T_i with $s_i = 0$. In this case, $\mathcal{E}\text{xe}_i(p_i, x_i) = \frac{\mathcal{E}\text{xe}_i(1, x_i)}{p_i} = \frac{\mathcal{E}\text{xe}_i^{\text{seq}}(x_i)}{p_i}$. The co-scheduling problem for such applications is denoted CoSchedCachePP.

4. Complexity Results

In this section, we focus on the CoSchedCache problem with rational numbers of processors in order to study the intrinsic complexity of co-scheduling with cache partitioning. We first prove that in an optimal execution, all applications must complete at the same time when using rational numbers of processors (Section 4.1). We remind that CoSchedCache is NP-complete, even for perfectly parallel applications (Section 4.2), and we show several dominance results on the optimal solution (Section 4.3). While some of these dominance results only hold for perfectly parallel applications, they will guide the design of heuristics for general applications in Section 5.

4.1. All applications complete at the same time

Lemma 1. *To minimize the makespan when using rational numbers of processors, all applications must finish at the same time.*

Proof. Consider n applications $T_1, \ldots, T_n$ that obey Amdahl's law, and a solution $\mathcal{S} = \{(p_i, x_i)\}_{1 \leq i \leq n}$ to CoSchedCache. Let $D_{\mathcal{S}} = \max_i \mathcal{E}\text{xe}_i(p_i, x_i)$ be the makespan of this solution. For simplicity, we let

$$\begin{aligned} A_i &= 1 + f_i \left(l_s + l_l \cdot \min \left(1, \frac{m^i_{\text{1MB}\mathcal{S}_s}}{\left(\frac{x_i C_s}{10^6}\right)^{\alpha}} \right) \right), \\ b_i &= A_i w_i s_i, \\ c_i &= A_i w_i (1 - s_i) \end{aligned}$$

Hence, $\mathcal{E}\text{xe}_i(p_i, x_i) = b_i + \frac{c_i}{p_i}$. The set of applications whose execution time is exactly $D_{\mathcal{S}}$ is denoted by $I_{\mathcal{S}}$.

We show the result by contradiction. We consider an optimal solution $\mathcal{S}$ whose subset $I_{\mathcal{S}}$ has minimal size (i.e., for any other optimal solution $\mathcal{S}_o$, $|I_{\mathcal{S}}| \leq |I_{\mathcal{S}_o}|$). Then we show that if $|I_{\mathcal{S}}| \neq n$, we can construct a solution $\mathcal{S}'$ with either (i) a smaller makespan if $|I_{\mathcal{S}}| = 1$ (contradicting the optimality hypothesis), or (ii) one less application whose execution time is exactly $D_{\mathcal{S}}$ (contradicting the minimality hypothesis).

Assume $|I_{\mathcal{S}}| \neq n$, let $T_{i_0} \in I_{\mathcal{S}}$ and $T_{i_1} \notin I_{\mathcal{S}}$. We have $\mathcal{E}\text{xe}_{i_1}(p_{i_1}, x_{i_1}) < \mathcal{E}\text{xe}_{i_0}(p_{i_0}, x_{i_0}) = D_{\mathcal{S}}$, that is

$$b_{i_1} + \frac{c_{i_1}}{p_{i_1}} < b_{i_0} + \frac{c_{i_0}}{p_{i_0}}, \text{ and hence } \quad (b_{i_1} - b_{i_0}) p_{i_0} p_{i_1} - c_{i_0} p_{i_1} + c_{i_1} p_{i_0} < 0. \tag{5}$$

We now prove that we can always find $0 < \varepsilon < p_{i_1}$ s.t. $\mathcal{E}\text{xe}_{i_0}(p_{i_0}, x_{i_0}) > \mathcal{E}\text{xe}_{i_0}(p_{i_0} + \varepsilon, x_{i_0}) > \mathcal{E}\text{xe}_{i_1}(p_{i_1} - \varepsilon, x_{i_1})$, i.e.,

$$D_{\mathcal{S}} = b_{i_0} + \frac{c_{i_0}}{p_{i_0}} > b_{i_0} + \frac{c_{i_0}}{p_{i_0} + \varepsilon} > b_{i_1} + \frac{c_{i_1}}{p_{i_1} - \varepsilon} \ .$$

The left part of inequality $b_{i_0} + \frac{c_{i_0}}{p_{i_0}} > b_{i_0} + \frac{c_{i_0}}{p_{i_0} + \varepsilon}$ is always true when $\varepsilon > 0$. For the right part of inequality above, we have:

$$-(b_{i_1} - b_{i_0})\varepsilon^2 + [(p_{i_1} - p_{i_0})(b_{i_1} - b_{i_0}) + c_{i_0} + c_{i_1}]\varepsilon + (b_{i_1} - b_{i_0}) p_{i_0} p_{i_1} - c_{i_0} p_{i_1} + c_{i_1} p_{i_0} < 0. \tag{6}$$

From Equation (5), we know that $(b_{i_1} - b_{i_0}) p_{i_0} p_{i_1} - c_{i_0} p_{i_1} + c_{i_1} p_{i_0} < 0$, so we can always find a $0 < \varepsilon < p_{i_1}$ that could make Equation (6) satisfied.

Then clearly, $\mathcal{S}' = \{(p'_i, x_i)\}_i$ where p'_i is (i) p_i if $i \notin \{i_0, i_1\}$, (ii) $p_{i_0} + \varepsilon$ if $i = i_0$, (iii) $p_{i_1} - \varepsilon$ if $i = i_1$, is a valid solution: we have the property $\sum_i p'_i = \sum_i p_i \leq p$, and $\sum_i x'_i = \sum_i x_i \leq 1$.

Hence,

- If $|I_{\mathcal{S}}| = 1$, then for all i, $\mathcal{E}\text{xe}_i(p'_i, x_i) < D_{\mathcal{S}}$, hence showing that $\mathcal{S}$ is not optimal;
- Else, $I_{\mathcal{S}'} = I_{\mathcal{S}} \setminus \{i_0\}$, and $D_{\mathcal{S}'} = D_{\mathcal{S}}$, hence showing that $\mathcal{S}$ is not minimal.

This shows that necessarily, $|I_{\mathcal{S}}| = n$. □

4.2. Intractability

We have shown in [ABD+17] that the problem is NP-complete, even for perfectly parallel applications.

Definition 2 (CoSchedCachePP-Dec)**.** *Given n perfectly parallel applications $T_1, \ldots, T_n$ and a platform with p identical processors sharing a cache of size C_s, and given a bound K on the makespan, does there exist a schedule $\{(p_1, x_1), \ldots, (p_n, x_n)\}$, where p_i and x_i are nonnegative rational numbers with $\sum_{i=1}^{n} p_i \leq p$ and $\sum_{i=1}^{n} x_i \leq 1$, such that $\max_{1 \leq i \leq n} \mathcal{E}xe_i(p_i, x_i) \leq K$?*

The proof of intractability is done thanks to a reformulation of the problem using the following Lemma:

Lemma 2. CoSchedCachePP *can be rewritten as finding the optimal cache partitioning strategy $\mathcal{X} = \{x_1, \ldots, x_n\}$ that minimizes the completion time of an optimal solution:*

$$\frac{1}{p} \sum_{i=1}^{n} \mathcal{E}xe_i(1, x_i). \tag{7}$$

Theorem 1. CoSchedCachePP-Dec *is NP-complete.*

4.3. Dominance results for perfectly parallel applications

In this section, we provide dominance results that will guide the design of heuristics. The dominance results are for perfectly parallel applications ($s_i = 0$) but we give intuition on how to extend this work for Amdahl applications in Section 4.4. Finally, we further assume that application memory footprints are larger than the cache size ($a_i = +\infty$), and we assume rational numbers of processors.

The core of the previous intractability result relies on the hardness to determine the set of applications that receive a cache fraction (denoted by I_C) and those that do not (denoted by $\overline{I_C}$). In this section, we show (i) how to determine the optimal solution when these sets I_C and $\overline{I_C}$ are known, and (ii) whether one can disqualify some partitions as being sub-optimal.

In particular, we define a set of partitions of applications that we call dominant (Definition 4). We show that (i) if a partition of applications $I_C, \overline{I_C}$ is dominant, then we can compute the minimum execution time for this partition, and (ii) if a partition is not dominant, then we can find a better dominant partition. We start by rewriting the problem when the partitioning $I_C, \overline{I_C}$ of applications is known:

Definition 3 (CSCPP-Part$\left(I_C, \overline{I_C}\right)$)**.** *Given a set of applications $T_1, \ldots, T_n$ and a partition $I_C, \overline{I_C}$, the problem* CSCPP-Part$\left(I_C, \overline{I_C}\right)$ *(for* CoSchedCachePP-Part*) is to find a set $\mathcal{X} = \{x_1, \ldots, x_n\}$ that minimizes the execution time:*

$$\frac{1}{p} \left(\sum_{i \in \overline{I_C}} w_i (1 + f_i (l_s + l_l)) + \sum_{i \in I_C} w_i (1 + f_i l_s + f_i l_l \frac{d_i}{x_i^{\alpha}}) \right)$$

under the constraints $x_i = 0$ if $i \in \overline{I_C}$, $x_i > d_i^{1/\alpha}$ if $i \in I_C$, and $\sum_{1 \leq i \leq n} x_i \leq 1$.

We now relax some bounds in CSCPP-Part$\left(I_C, \overline{I_C}\right)$ and define CSCPP-Ext$\left(I_C, \overline{I_C}\right)$, which is the same problem except that the constraints on the x_i's when $i \in I_C$ is relaxed: we have instead $x_i \geq 0 \quad$ if $i \in I_C$.

A solution of CSCPP-Part$\left(I_C, \overline{I_C}\right)$ is a solution of CSCPP-Ext$\left(I_C, \overline{I_C}\right)$, because we simply removed the constraints $x_i > d_i^{1/\alpha}$ in the latter problem. Hence the execution time of the optimal solution of CSCPP-Ext$\left(I_C, \overline{I_C}\right)$ is lower than that of CSCPP-Part$\left(I_C, \overline{I_C}\right)$.

Furthermore, given a solution of CSCPP-Ext$\left(I_C, \overline{I_C}\right)$, one can easily see that its execution time in CoSchedCache will be lower (the objective function is lower since it involves a minimum for all applications in I_C).

Lemma 3. *Given a set of applications $T_1, \ldots, T_n$ and a partition $I_C, \overline{I_C}$, the optimal solution to* CSCPP-Ext$\left(I_C, \overline{I_C}\right)$ *is*

$$x_i = \frac{(w_i f_i d_i)^{1/(\alpha+1)}}{\sum_{j \in I_C} (w_j f_j d_j)^{1/(\alpha+1)}} \qquad \text{if } i \in I_C,$$

$$x_i = 0 \qquad \text{otherwise.}$$

The proof is available in the companion research report [ABD+17].

Definition 4 (Dominant partition)**.** *Given a set of applications $T_1, \ldots, T_n$, we say that a partition of these applications $I_C, \overline{I_C}$ is* dominant*, if for all $i \in I_C$,*

$$\frac{(w_i f_i d_i)^{1/(\alpha+1)}}{\sum_{j \in I_C} (w_j f_j d_j)^{1/(\alpha+1)}} > d_i^{1/\alpha}.$$

We can now state the following result:

Theorem 2. *If a partition $I_C, \overline{I_C}$ is not dominant, then we can compute in polynomial time a better solution.*

The proof is available in the companion research report [ABD+17].

We can show a second dominance result characterizing the optimal solution:

Theorem 3. *If a partition $I_C, \overline{I_C}$ is dominant, then the optimal solution to* CSCPP-Part$\left(I_C, \overline{I_C}\right)$ *is:*

$$x_i = \frac{(w_i f_i d_i)^{1/(\alpha+1)}}{\sum_{j \in I_C} (w_j f_j d_j)^{1/(\alpha+1)}} \qquad \text{if } i \in I_C;$$
$$x_i = 0 \qquad \text{otherwise.}$$

Proof. This is a corollary of Lemma 3.

Indeed, this solution is the optimal solution to CSCPP-Ext$\left(I_C, \overline{I_C}\right)$ and it is a valid solution to CSCPP-Part$\left(I_C, \overline{I_C}\right)$, hence it is the optimal solution to CSCPP-Part$\left(I_C, \overline{I_C}\right)$. □

4.4. Extension of the dominance criterion for Amdahl applications

Finally, we provide extended definitions for non-perfectly parallel applications, by defining the dominant partition of both the parallel part and the sequential part of such applications.

Definition 5 (Dominant partition of parallel part)**.** *Given a set of applications $T_1, \ldots, T_n$, we say that a partition of these applications $I_C, \overline{I_C}$ is* dominant *for the parallel part if for all $i \in I_C$,*

$$\frac{(w_i f_i d_i (1-s_i))^{1/(\alpha+1)}}{\sum_{j \in I_C} (w_j f_j d_j (1-s_j))^{1/(\alpha+1)}} > d_i^{1/\alpha}.$$

Definition 6 (Dominant partition of sequential part)**.** *Given a set of applications $T_1, \ldots, T_n$, we say that a partition of these applications $I_C, \overline{I_C}$ is* dominant *for the sequential part if for all $i \in I_C$,*

$$\frac{(w_i f_i d_i s_i)^{1/(\alpha+1)}}{\sum_{j \in I_C} (w_j f_j d_j s_j)^{1/(\alpha+1)}} > d_i^{1/\alpha}.$$

The intuition behind these two definitions is the following: recall from Lemma 1 that the execution time is defined as $\mathcal{E}\text{xe}_i(p_i, x_i) = b_i + \frac{c_i}{p_i}$, with

$$A_i = 1 + f_i \left(l_s + l_l \cdot \min \left(1, \frac{m^i_{1\text{MB}\mathcal{S}_s}}{\left(\frac{x_i C_s}{10^6}\right)^{\alpha}} \right) \right),$$
$$b_i = A_i w_i s_i,$$
$$c_i = A_i w_i (1 - s_i).$$

We can observe that s_i, the sequential fraction, is key to decide which parts b_i or $\frac{c_i}{p_i}$ we should favor to minimize $\mathcal{E}\text{xe}_i(p_i, x_i)$. If $s_i << \frac{1}{p_i}$, then $\frac{c_i}{p_i}$ dominates the execution time, i.e., $\mathcal{E}\text{xe}_i(p_i, x_i) \approx c_i$. Hence the application could be seen as a perfectly parallel application where the new number of computing operations to do is $\tilde{w_i} = w_i(1 - s_i)$. Then Definition 5 is just a consequence of applying the definition of Dominant Partition to this new application.

Symmetrically, if s_i is large in front of one over the number of processors assigned to an application, then b_i dominates the execution time. Intuitively in this case, the number of processors by application is less important (and we will have a fair balance of processors). Hence, we want to favor applications with large values of $s_i w_i f_i d_i$.

We verify these intuitions experimentally in Section 6.

5. Heuristics

In this section, we aim at designing efficient heuristics for general applications that obey Amdahl's law, and whose memory footprints are larger than the cache size ($a_i = +\infty$). However, the CoSchedCache problem seems to be very difficult for such applications, as seen in Section 4.

We first explain how heuristics work, in particular to assign (rational numbers of) processors, in Section 5.1. The core of the heuristic consists in building a dominant partition, and we detail different possibilities to do so in Section 5.2. Finally, we propose a way to round the number of processors in case we need an integer number of processors, for instance if no multi-threading is allowed (see Section 5.3).

5.1. Structure of heuristics

We simplify the design of the heuristics by temporarily allocating processors as if the applications were perfectly parallel, and then concentrating on strategies that partition the cache efficiently among some applications (and give no cache fraction to remaining ones). In accordance with Theorem 2, our goal is to compute dominant partitions. Recall that I_C represents the subset of applications that receive a fraction of the cache. Once a dominant partition is given, we obtain the schedule $\mathcal{S} = \{(x_i, p_i)\}_i$ as follows: first we determine the x_i's with Theorem 3, and then we recompute the p_i's so that all applications complete simultaneously at time K. Indeed, while Lemma 2 does not hold for Amdahl applications, we still know thanks to Lemma 1 that all applications should complete simultaneously.

However, there is no longer a nice analytical characterization of the makespan K, hence we use a binary search to compute K as follows: for each application T_i, the execution time writes $(s_i + \frac{1-s_i}{p_i})c_i = K$, where s_i is the sequential fraction, and $c_i = w_i(1 + f_i(l_s + l_l \frac{d_i}{x_i^\alpha}))$ if $T_i \in I_C$, or $c_i = w_i(1 + f_i(l_s + l_l))$ otherwise. From $\sum_{i=1}^n p_i = p$, we derive the equation

$$\sum_{i=1}^{n} \frac{1 - s_i}{\frac{K}{c_i} - s_i} = p$$

and we compute K through a binary search. A lower (resp. upper) bound for K is to assign p (resp. 1) processor(s) to each application.

5.2. Computing a dominant partition

To compute dominant partitions, we use two greedy strategies:

- Dom: we start with $I_C = \mathcal{I}$ and greedily remove some applications from I_C until we have a dominant partition (see Algorithm 1); NotDom(i, I_C) returns true if i does not satisfy the definition of dominant partition for I_C;
- DRev: initially I_C is empty, and we greedily add applications while I_C remains dominant (see Algorithm 2); IsDom(I'_C) returns true if I'_C is a dominant partition.

Both strategies come in three flavors, depending on the dominance definition that we use. From Definition 4, we get that NotDom(i, I_C) is true if and only if $\frac{(w_i f_i d_i)^{1/(\alpha+1)}}{d_i^{1/\alpha}} \leq \sum_{j \in I_C} (w_j f_j d_j)^{1/(\alpha+1)}$, and IsDom$(I'_C)$ is true if and only if $\forall i \in I'_C$, $\frac{(w_i f_i d_i)^{1/(\alpha+1)}}{d_i^{1/\alpha}} > \sum_{j \in I'_C} (w_j f_j d_j)^{1/(\alpha+1)}$ (strategies Dom and DRev). If we use Definition 6, we simply replace all w_k's by $w_k s_k$ (strategies DomS and DRevS focusing on the sequential part), while with Definition 5, we replace all w_k's by $w_k(1 - s_k)$ (strategies DomP and DRevP focusing on the parallel part).

Algorithm 1: DOM strategy, starting with all applications

```
1  procedure DOM (I, choice) begin
2      I_C ← I;
3      while ∃i ∈ I_C s.t. NOTDOM(i, I_C) do
4          k ← choice(I_C);
5          I_C ← I_C \ {k};
6          if I_C = ∅ then break;
7      end
8      Ī_C ← I \ I_C;
9      return (I_C, Ī_C);
10 end
```

Algorithm 2: DREV strategy, starting from empty set

```
1  procedure DREV (I, choice) begin
2      Ī_C ← I; I_C ← ∅;
3      k ← choice(Ī_C);
4      I'_C ← {k};
5      while ISDOM(I'_C) do
6          I_C ← I'_C;
7          Ī_C ← Ī_C \ {k};
8          if Ī_C = ∅ then break;
9          k ← choice(Ī_C);
10         I'_C ← I'_C ∪ {k};
11     end
12     return (I_C, Ī_C);
13 end
```

Figure 1: Two strategies to build dominant partitions.

For each of these strategies, the greedy criterion to select the next application is the *choice* function taken from the following three alternatives:

- RANDOM: $choice(\mathcal{I})$ picks up randomly one application among all applications;
- MINRATIO considers the ratio that appears in Definition 4, 6 or 5 (dominant partitions), and chooses an application with a small ratio; for DOM and DREV, we have:

$$choice(\mathcal{I}) = \arg\min_{i \in \mathcal{I}} \left(\frac{(w_i f_i d_i)^{1/(\alpha+1)}}{d_i^{1/\alpha}} \right);$$

  and we replace w_i by $w_i s_i$ in DOMS and DREVS, or by $w_i(1-s_i)$ in DOMP and DREVP;
- MAXRATIO proceeds the other way round, by choosing an application with a large ratio, simply replacing the arg min by an arg max.

The intuition behind these heuristics is the following: applications that make the solution non dominant for DOM and DREV are such that (see Definition 4):

$$\frac{(w_i f_i d_i)^{1/(\alpha+1)}}{d_i^{1/\alpha}} \leq \sum_{j \in I_C} (w_j f_j d_j)^{1/(\alpha+1)}.$$

Hence, we expect to reach dominance faster by removing from a non-dominant solution applications with low $\frac{(w_i f_i d_i)^{1/(\alpha+1)}}{d_i^{1/\alpha}}$ (left term of the equation). Intuitively, DOM, DOMS and DOMP should work well with the MINRATIO criterion. For symmetric reasons, we expect DREV, DREVS and DREVP to work well with the MAXRATIO criterion. These intuitions will be experimentally confirmed in Section 6.

Altogether, by combining six strategies, and with three different *choice* functions for each strategy, we obtain 18 heuristics to build dominant partitions. We denote by DOM-MINRATIO the DOM strategy using MINRATIO as a *choice* function, and we use a similar notation for all heuristics.

5.3. Integer processor assignment

Based on the rational cache allocation, we want to give an integer processor allocation in order to tackle architectures that do not allow to share processors between applications through multi-threading. The choice functions above are first used to build a dominant partition, then we assign cache based on that partition to obtain the x_i's. In Algorithm 3, the set $\mathcal{I}$ contains all applications and x is the set that contains all x_i's. Finally, p is the total number of processors and n the total number of applications (i.e., $n = |\mathcal{I}|$). After the

App	Description
CG	Uses conjugate gradients method to solve a large sparse symmetric positive definite system of linear equations
BT	Solves multiple, independent systems of block tridiagonal equations with a predefined block size
LU	Solves regular sparse upper and lower triangular systems
SP	Solves multiple, independent systems of scalar pentadiagonal equations
MG	Performs a multi-grid solve on a sequence of meshes
FT	Performs discrete 3D fast Fourier Transform

Figure 2: Description of the NPB benchmarks.

App	w_i	f_i	$m^i_{40\text{MB}\mathcal{S}_s}$
CG	5.70E+10	5.35E-01	6.59E-04
BT	2.10E+11	8.29E-01	7.31E-03
LU	1.52E+11	7.50E-01	1.51E-03
SP	1.38E+11	7.62E-01	1.51E-02
MG	1.23E+10	5.40E-01	2.62E-02
FT	1.65E+10	5.82E-01	1.78E-02

Figure 3: Experimental values from NPB benchmarks.

cache is assigned, we initialize processor assignment by giving one processor to each application, and the remaining processors are assigned in a greedy way: assign one processor to the application currently with longest execution time, until all processors are assigned. It should be noted that integer processor assignment will only work when $p \geq n$, since each application needs at least one processor.

Algorithm 3: Integer processor assignment

```
1  procedure IntegerProcessor (x, p, I)
2  begin
3      for i ∈ I do p'_i = 1;
4      p_remain = p − n;
5      while p_remain > 0 do
6          i = arg max_{k∈I} (Exe_k(p'_k, x_k));
7          p'_i = p'_i + 1;
8          p_remain = p_remain − 1;
9      end
10     return p'_i;
11 end
```

6. Simulations

To assess the efficiency of the heuristics defined in Section 5, we have performed extensive simulations. The simulation settings are discussed in Section 6.1, and results are presented in Section 6.2 (comparison of the 18 heuristics of Section 5), Section 6.3 (assessing the gain due to co-scheduling), and Section 6.4 (with integer numbers of processors). The code is publicly available at `http://perso.ens-lyon.fr/loic.pottier/archives/cache-int.zip`.

6.1. Simulation settings

We use data from applicative benchmarks to run the experiments. Figure 2 provides a brief description of the NAS Parallel Benchmark (NPB) suite [BBB+91], and Figure 3 shows the parameters for these six HPC applications. We obtain the values shown in Figure 3 by instrumenting and simulating the benchmarks (*CLASS=A*) on 16 cores using PEBIL [LTCS10]. For the simulations, we use a cache configuration representing an Intel Xeon CPU E5-2690, with a 40MB last level cache per processor of 8 cores. Since the cache miss ratio is defined for a 40MB cache, we have $d_i = m^i_{40\text{MB}\mathcal{S}_s} \left(\frac{40\times 10^6}{C_s}\right)^{\alpha}$.

To build a set of n applications, we pick randomly n times one application among the six applications defined by Table 3, the number of application wanted. In additions, for each of these n applications, the work w_i is randomly taken between 1E+8 and 1E+12. Other data sets building upon these applications have

been used (see the companion research report [ABD+17]), and the results are very similar. The sequential fraction of work s_i is taken randomly between 1% and 15%.

For the execution platform, we consider one manycore *Sunway TaihuLight* [Don16] with 256 processors and a shared memory of 32GB. We chose this platform because of its high core count. Strictly speaking, this platform does not have a last level cache (LLC), but the shared memory can be seen as the LLC, using the disk as the large memory. We have $C_s = 32 \times 10^9$. The large storage latency l_l is set to 1. The small storage latency l_s is set to 0.17. According to the literature [KKSM13, MHSN15, PB14], the last level cache (LLC) latency is on average four to ten times better than the DDR latency, and we enforce a ratio of 5.88 in the simulations. We have used different ratios in [ABD+17], and they lead to similar results. Finally, the Power Law parameter is set to $\alpha = 0.5$. We execute each heuristic 50 times and we compute the average *makespan*, i.e., the longest execution time among all co-scheduled applications.

6.2. Comparison of the heuristics

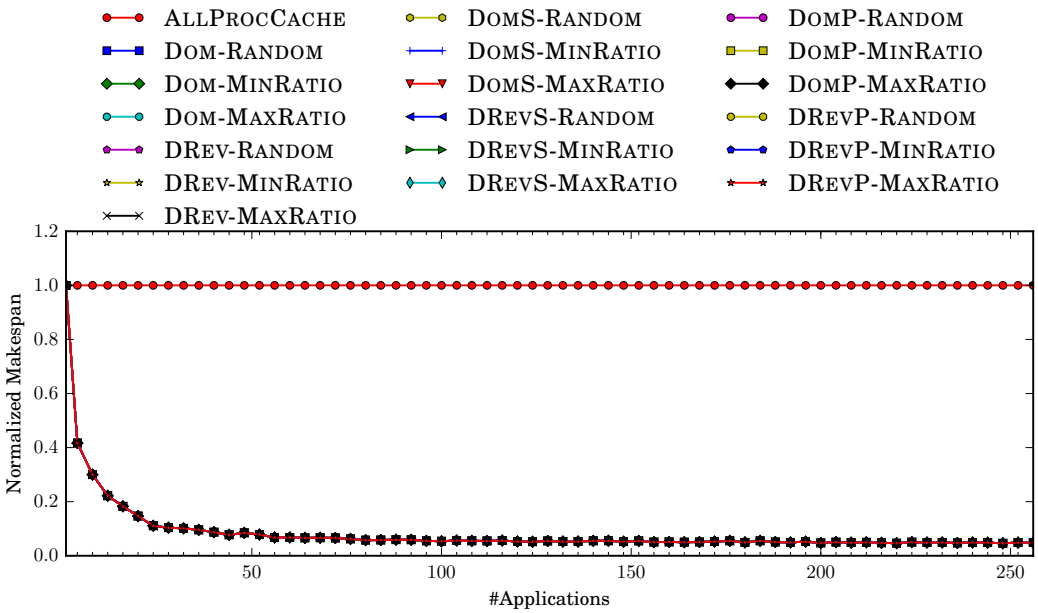


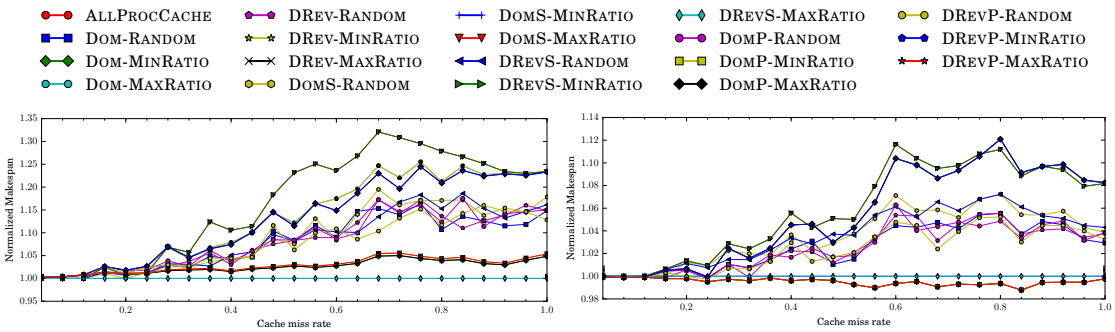
Figure 4: Comparison of all dominant partition heuristics on 256 processors.

Figure 4 shows the normalized makespan obtained by all of the heuristics building dominant partitions. We set the number of processors to 256. Results are normalized with the makespan of AllProcCache, which is the execution without any co-scheduling: in the AllProcCache heuristic, applications are executed sequentially, each using all processors and all the cache. We vary the number of applications between 1 and 256. The eighteen heuristics obtain similarly good results, with a gain of 85% over AllProcCache as soon as there are at least 50 applications.

Since all eighteen variants show the same performance on the previous data sets, we investigate the impact of the cache miss rate by varying it between 0 and 1 with a LLC of $C_s = 1GB$ in Figure 5. Results are now normalized with DomS-MinRatio in both figures, which enables to zoom out the differences.

The first noticable result from Figure 5 is that for all versions of the strategies that build dominant strategies, MinRatio performs better with strategies that remove applications from the I_C (Dom, DomS, DomP), whereas MaxRatio works better with strategies that add applications to the I_C (DRev, DRevS, DRevP). This confirms the mathematical intuition presented in Section 5.

Furthermore, we confirm the mathematical intuition on the influence of the Amdahl factor (s_i) presented in Section 4.4:



(a) s_i randomly set between 0.01 and 0.15.

(b) s_i randomly set between 0.001 and 0.01.

Figure 5: Impact of the cache miss ratio $m^i_{40\text{MB}\mathcal{S}_s}$ with a 1GB cache and 16 applications.

- We observe that in Figure 5a, when the sequential fraction is not negligible (s_i chosen uniformly at random between 0.01 and 0.15), DomS-MinRatio and DRevS-MaxRatio are always the best (their plots overlap), with a gain from 10 to 15% with respect to the random-based heuristics when the cache miss rate is greater than 0.5.
- On the contrary, when it is negligible (s_i chosen uniformly at random between 0.001 and 0.01), then the DomP-MinRatio and DRevP-MaxRatio versions perform better.

Note that overall, the observable differences between heuristics is mainly when the cache miss ratio is large. According to current data, $m_{40\text{MB}\mathcal{S}_s}$ ranges from 1E-02 to 1E-04 (see Table 3). In addition, these differences are visible only with a small shared memory (1GB in the example), while our execution platform has a 32GB shared memory. Overall, for the system used in these simulations, all heuristics perform similarly, even though DomS-MinRatio and DRevS-MaxRatio seem to perform best in all other settings that we tried (see [ABD+17]).

In the following simulations, the sequential fraction will always, unless otherwise mentioned, be taken between 1% and 15%. Therefore, for clarity, we plot only one heuristic based on dominant partitions in the remaining simulations, namely DomS-MinRatio.

6.3. Gain with co-scheduling

In this section, we assess the gain due to co-scheduling by comparing DomS-MinRatio with AllProcCache and with three other heuristics:

- Fair gives $p_i = \frac{p}{n}$ processors, and a fraction of cache $x_i = \frac{f_i}{\sum_{j=1}^{n} f_j}$ to each application;
- 0cache gives no cache to any application, i.e., $x_i = 0$ for $1 \leq i \leq n$, and then it computes the p_i's so that all applications finish at the same time;
- RandomPart randomly partitions applications with and without cache. For those in cache, the x_i's are computed with the method used for dominant partitions. Then, the p_i's are computed so that all applications finish at the same time.

Impact of the number of applications. Figure 6 (normalized with AllProcCache on the left) shows the impact of the number of applications when the number of processors is set to 256. We see that DomS-MinRatio outperforms the other heuristics, hence showing the efficiency of our approach based on dominant partitions. Results are also normalized with DomS-MinRatio (on the right), so that we can better observe the differences between co-scheduling heuristics. Fair exhibits good results only for a small number of applications, when all applications can fit into cache. Otherwise, the use of dominant partitions is much more efficient, as seen with RandomPart, or even 0cache that does not use cache but ensures that all applications finish at the same time. These results show the accuracy of the model and the benefits of using dominant partitions. Also, we note the importance of cache partitioning, since the difference between 0cache and DomS-MinRatio relies on cache allocation.

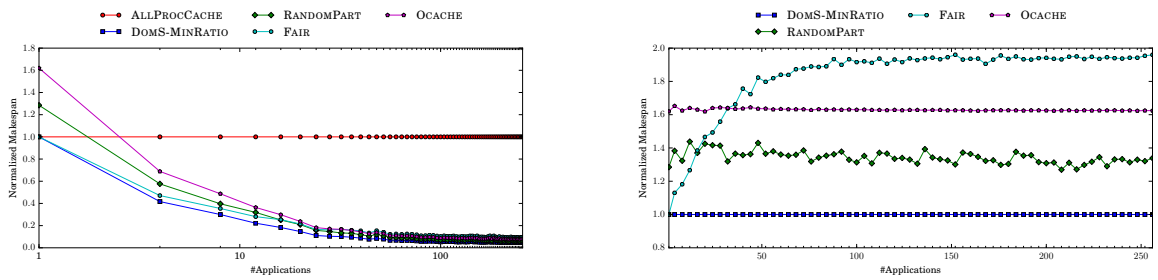


Figure 6: Impact of the number of applications.

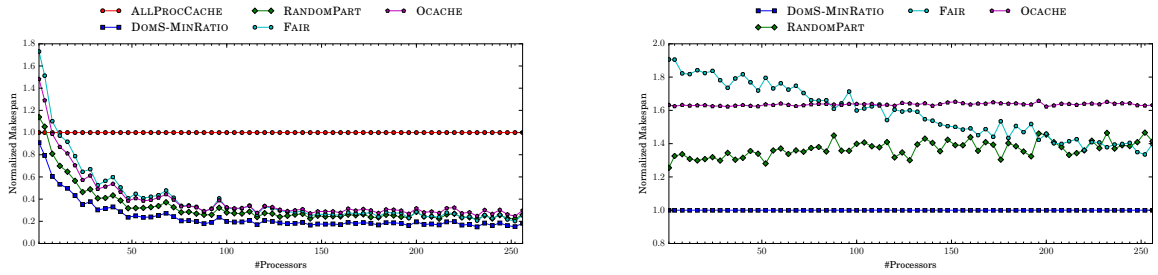


Figure 7: Impact of the number of processors.

Impact of the number of processors. Figure 7 (normalized with AllProcCache on the left) shows the impact of the number of processors when the number of applications is set to 16. When the number of processors increases, the gain of co-scheduling increases. In both figures, DomS-MinRatio and outperforms other methods. RandomPart, which builds a random partition instead of a dominant one, is outperformed by DomS-MinRatio, and the latter is the only heuristic that surpasses AllProcCache when the number of processors is low. So, building a dominant partition seems a good strategy to optimize the makespan.

The normalization with DomS-MinRatio (on the right) shows that when the number of processors increases, Fair becomes better, while RandomPart and 0cache are quite stable since they are based on the same model as DomS-MinRatio. The only difference between 0cache and DomS-MinRatio is the cache allocation strategy, and the gain from cleverly distributing cache fractions across applications exceeds 20%. With more applications, we obtain the same ranking of heuristics, except that Fair is always the worst heuristic: since there are less processors on average per application, a good co-scheduling policy is necessary (see [ABD+17] for detailed results).

Impact of the sequential fraction of work. Figure 8 (normalized with AllProcCache) shows the impact of the sequential part s_i when the number of processors is set to 256. The number of applications is set to 16. As expected, when the sequential fraction of work increases, all co-scheduling heuristics perform better than AllProcCache, and DomS-MinRatio is always the best heuristic. It leads to a gain of more than 50% when $s_i = 0.01$.

The normalization with DomS-MinRatio better shows the impact of the sequential part: we observe that when the sequential fraction of work increases, Fair obtains results closer to DomS-MinRatio.

Processor and cache repartition. Figure 9 shows the processor repartition and cache repartition when we vary the number of applications from 1 to 256 with 256 processors. We use an error bar plot where the error interval represents here the maximum and minimum number of processors (or cache fraction) allocated to an application. As expected, we observe that the range between minimum and maximum decreases when the number of applications increases. The processor allocation of Fair is not interesting, the maximum is

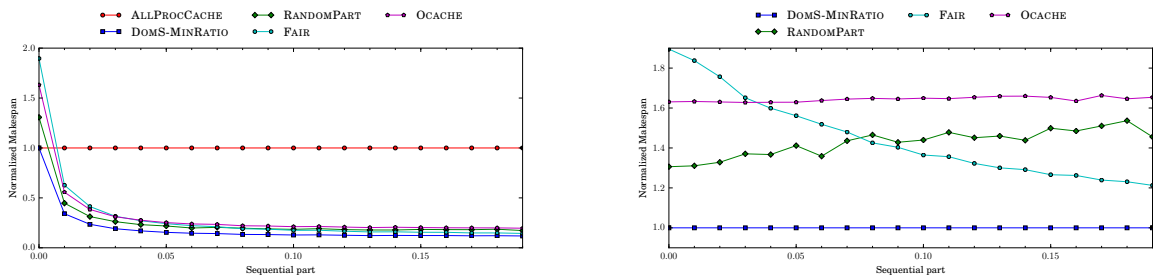


Figure 8: Impact of sequential fraction of work.

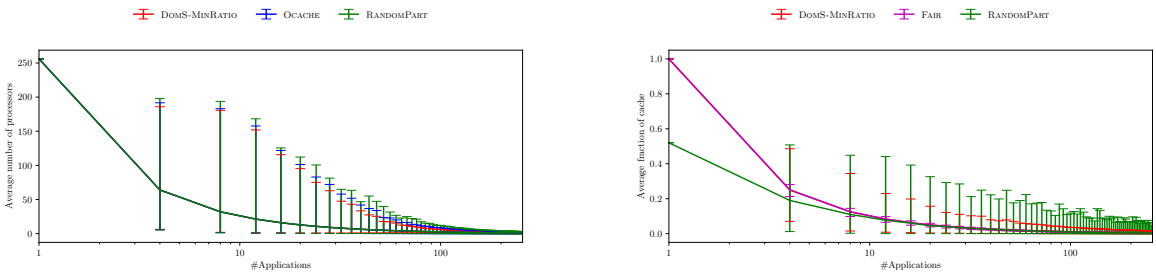


Figure 9: Processor and cache repartition with 256 processors.

always equal to the minimum because we allocate the same number of processors to each application.

Since all dominant partition heuristics give the same results, we only use DomS-MinRatio. The repartition of processors for 0Cache is interesting: it turns out to be very close to the repartition obtained with DomS-MinRatio, even though it is not using cache.

Summary. To summarize, all heuristics based on dominant partitions are very efficient, especially when compared to the classical heuristics Fair (which shares the cache fairly between applications) and AllProcCache (which does no co-scheduling). The unexpected result that can be observed is that the gain brought by our heuristics comes even with very low sequential time (below 0.01)! This is unexpected since the natural intuition would be a behavior such as the one observed on Fair: a makespan up to 1.9 times longer than AllProcCache with low sequential time.

We show that the ratio processors/applications has a significant impact on performance: when many processors are available for a few applications, it is less crucial to use efficient cache-partitioning and all applications can share the cache, hence Fair obtains good results, close to DomS-MinRatio. Otherwise, RandomPart is the second best heuristic. A surprising information that also confirms the strength of our partition based heuristics is that *natural* heuristics such as Fair and AllProcCache perform worse than 0Cache our implementation with no usage of cache.

All heuristics run within a very small time (less than ten seconds in the worst of the settings used, to be compared with a typical application execution time in hours or days), hence they can be used in practice with a very light overhead.

6.4. With an integer number of processors

In this section, we study the impact of rounding the number of processors to an integer number on heuristics. We focus again mainly on DomS-MinRatio, and we add the suffix Int to heuristic names to denote the fact that we use Algorithm 3 to compute an integer processor allocation.

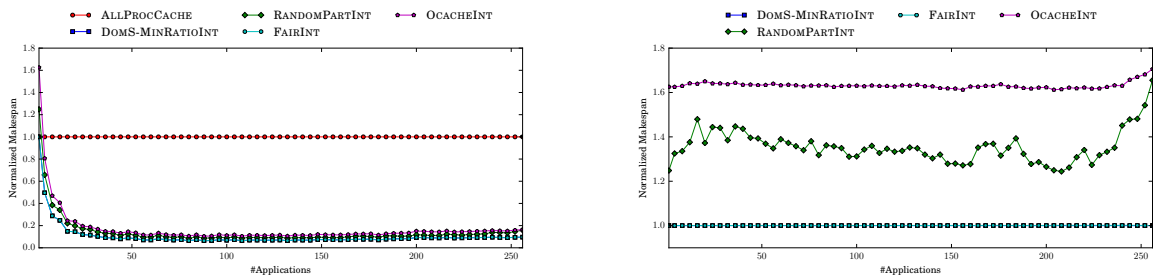


Figure 10: Impact of the number of applications.

Impact of the number of applications. In this simulation, we vary the number of applications from 1 to 256 on 256 processors. Figure 10 is normalized with AllProcCache (on the left), and heuristics obtain a similar relative performance as in Section 6.3, with a gain of 90% over AllProcCache as soon as there are at least 50 applications. The right side of Figure 10 shows the performance of the same heuristics but normalized with DomS-MinRatioInt. As expected, 0cacheInt is the worst, and RandomPartInt performs always in the middle between 0cacheInt and FairInt. As we use the same algorithm to round the rational processor allocation, the differences in performance mostly rely on cache allocation.

The fact that FairInt and DomS-MinRatioInt give similar results show that the cache allocation of DomS-MinRatioInt must not be far from the fair distribution of FairInt. However, contrarily to Fair, processors are not equally shared between applications but distributed according to their needs, hence the much better performance of FairInt compared to Fair.

Simulations showing the impact of the number of processors and of the sequential fraction of work give similar results, with FairInt and DomS-MinRatioInt overlapping and beating other heuristics. We refer to the companion research report for details [ABD+17].

Impact of the sequential fraction and the cache miss rate. As DomS-MinRatioInt and FairInt show the same performance, we study the impact of the sequential fraction and the cache miss rate, as we did in Section 6.2, in Figure 11. The number of applications is set to 16 and the number of processors to 256 with a LLC of $C_s = 1GB$. The results are normalized with DomS-MinRatioInt. In the left figure, we compare all dominant partition heuristics by varying the sequential fraction when the cache miss rate is set to 0.8 in order to see differences between heuristics. We note that the dominant partition heuristics favoring the sequential part outperform the others, especially the ones favoring the parallel part. Dom-MinRatioInt and DRev-MaxRatioInt overlap with DomS-MinRatioInt. All variants using Random criterion perform on average around 1.10. As expected, giving more cache to applications with bigger sequential fractions is better. In the right figure, we vary the cache miss rate between 0 and 1 This figure is interesting due to the difference of performance between DomS-MinRatioInt and FairInt. Clearly, the difference of performance between heuristics when we use integer processors rely on cache allocation. When the cache miss ratio increases, the performance of DomS-MinRatioInt becomes better. When the cache miss rate is larger than 0.01, DomS-MinRatioInt outperforms all other heuristics, and we obtain an average gain of 10% on FairInt. The performance of 0cacheInt becomes better when the cache miss rate increases.

Summary. To summarize, when we use integer processors, all heuristics based on dominant partitions are still very efficient, but those that favor either the sequential part or none of them perform better. The main difference between results with rational and integer processor assignments is that DomS-MinRatioInt and FairInt overlap if the cache miss rate is low (less than 1%), because of the better processor assignment for FairInt. We show that the cache miss rate has a significant impact on performance: when many cache misses occur, it is more crucial to use efficient cache-partitioning and all applications can share the cache, hence DomS-MinRatioInt outperforms FairInt when the cache miss rate is larger than 10%. As expected, DomS-MinRatioInt performs better when the cache miss rate increases. Otherwise, RandomPartInt is the third best heuristic, followed by 0cacheInt that does not use the cache.

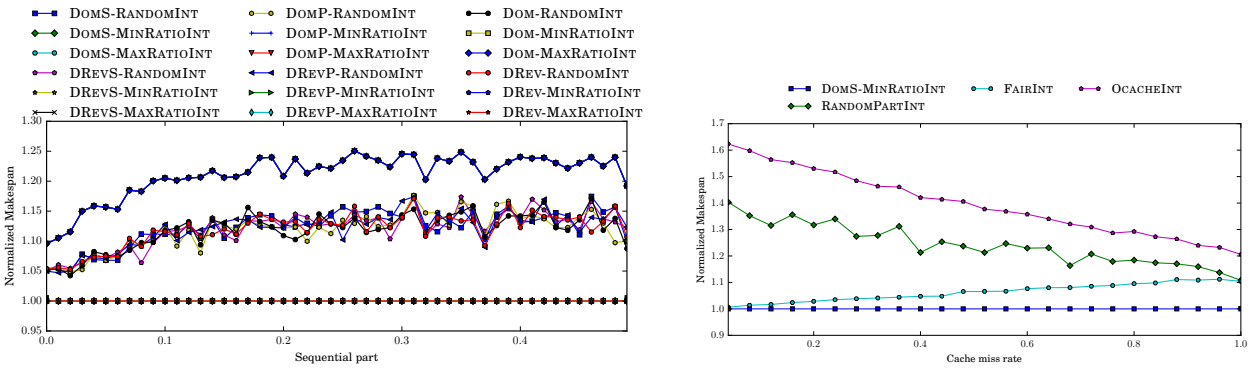


Figure 11: Impact of the sequential fraction and the cache miss rate.

7. Conclusion

In this paper, we have provided a preliminary study on co-scheduling algorithms for cache-partitioned systems, building upon a theoretical study. The two key scheduling questions are (i) which proportion of cache and (ii) how many processors should be given to each application. For rational numbers of processors, we proved that the problem is NP-complete, but we have been able to characterize optimal solutions for perfectly parallel applications by introducing the concept of *dominant partitions*: for such applications, we have computed the optimal proportion of cache to give to each application in the partition. Furthermore, we have provided explicit formulas to express the number of processors to assign to each application.

Several polynomial-time heuristics focusing on Amdahl's applications have been built upon these results, both for rational and integer numbers of processors. Extensive simulation results demonstrate that the use of dominant partitions always leads to better results than more naive approaches, as soon as there is a small sequential fraction of work in application speedup profiles. The concept of sharing the cache only between a subset of applications seems highly relevant, since even an approach with a random selection of applications that share the cache leads to good results. Also, a clever partitioning of the cache pays off quite well, since our heuristics lead to a significant gain compared to an approach where no cache is given to applications. Overall, the heuristics appear to be very useful for general applications, even though their cache allocation strategy rely mainly on simulating a perfectly parallel profile.

Future work will be devoted to gain access to, and conduct real experiments on, a cache-partitioned system with a high core count: this would allow us to further validate the accuracy of the model and to confirm the impact of our promising results. On the theoretical side, we plan to focus on the problem with integer numbers of processors and we hope to derive interesting results that could help design even more efficient heuristics.

Acknowledgments

The authors would like to thank the reviewers for their insightful comments. Yves Robert is with the Institut Universitaire de France. This research was possible thanks to an Inria grant and funding from Vanderbilt university.

[ABD+17] Guillaume Aupy, Anne Benoit, Sicheng Dai, Loic Pottier, Padma Raghavan, Yves Robert, and Manu Shantharam. Co-scheduling amdahl applications on cache-partitioned systems. Research report RR-9021, INRIA, 2017. Available at `graal.ens-lyon.fr/~abenoit`.

[Adv14] Advanced Scientific Computing Advisory Committee (ASCAC). Ten technical approaches to address the challenges of Exascale computing, 2014. `https://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20140210/Top10reportFEB14.pdf`.

[Amd67] G. Amdahl. The validity of the single processor approach to achieving large scale computing capabilities. In *AFIPS Conference Proceedings*, volume 30, pages 483–485. AFIPS Press, 1967.

[BBB+91] D. H. Bailey, E. Barszcz, J. T. Barton, D. S. Browning, R. L. Carter, L. Dagum, R. A. Fatoohi, P. O. Frederickson, T. A. Lasinski, R. S. Schreiber, H. D. Simon, V. Venkatakrishnan, and S. K. Weeratunga. The NAS Parallel Benchmarks – Summary and Preliminary Results. In *Proceedings of the 1991 ACM/IEEE Conference on Supercomputing*, SC'91, pages 158–165, New York, NY, USA, 1991. ACM.

[BCSM08] B. D. Bui, M. Caccamo, L. Sha, and J. Martinez. Impact of cache partitioning on multi-tasking real time embedded systems. In *4th IEEE Int. Conf. on Embedded and Real-Time Computing Systems and Applications*, pages 101–110. IEEE Computer Society, 2008.

[BZF10] Sergey Blagodurov, Sergey Zhuravlev, and Alexandra Fedorova. Contention-aware scheduling on multicore systems. *ACM Trans. Comput. Syst.*, 28(4):8:1–8:45, 2010.

[CE00] Franck Cappello and Daniel Etiemble. MPI Versus MPI+OpenMP on IBM SP for the NAS Benchmarks. In *SC '00*, Washington, DC, USA, 2000. IEEE Computer Society.

[DFB+12] Tyler Dwyer, Alexandra Fedorova, Sergey Blagodurov, Mark Roth, Fabien Gaud, and Jian Pei. A Practical Method for Estimating Performance Degradation on Multicore Processors, and Its Application to HPC Workloads. In *Proc. Int. conf. High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 83:1–83:11, 2012.

[DJF+15] D. Dauwe, E. Jonardi, R. Friese, S. Pasricha, A. A. Maciejewski, D. A. Bader, and H. J. Siegel. A methodology for co-location aware application performance modeling in multicore computing. In *Parallel and Distributed Processing Symposium Workshop (IPDPSW)*, pages 434–443. IEEE, 2015.

[Don16] Jack Dongarra. Report on the sunway taihulight system. *PDF). www. netlib. org. Retrieved June*, 20, 2016.

[GAB+15] Ana Gainaru, Guillaume Aupy, Anne Benoit, Franck Cappello, Yves Robert, and Marc Snir. Scheduling the I/O of HPC applications under congestion. In *IEEE Int. Parallel and Distributed Processing Symposium (IPDPS)*, pages 1013–1022, 2015.

[GSYY09] Nan Guan, Martin Stigge, Wang Yi, and Ge Yu. Cache-aware scheduling and analysis for multicores. In *Proc. 7th ACM Int. Conf. Embedded Software*, EMSOFT '09, pages 245–254. ACM, 2009.

[Hea15] Michael T Heath. A tale of two laws. *Int. J. High Performance Computing Applications*, 29(3):320–330, 2015.

[HSPE08] Allan Hartstein, Vijayalakshmi Srinivasan, T Puzak, and P Emma. On the nature of cache miss behavior: Is it $\sqrt{2}$. *The Journal of Instruction-Level Parallelism*, 10:1–22, 2008.

[HZJ16] L. He, H. Zhu, and S. A. Jarvis. Developing graph-based co-scheduling algorithms on multicore computers. *IEEE Trans. Parallel Distributed Systems*, 27(6):1617–1632, 2016.

[Int14] Intel. Intel 64 and IA-32 architectures software developer's manual. *Part 2*, 3B: System Programming Guide, 2014.

[JSCT08] Yunlian Jiang, Xipeng Shen, Jie Chen, and Rahul Tripathi. Analysis and approximation of optimal co-scheduling on chip multiprocessors. In *Proc. 17th Int. Conf. Parallel Architectures Compilation Techniques*, PACT '08, pages 220–229. ACM, 2008.

[KKSM13] E. Kultursay, M. Kandemir, A. Sivasubramaniam, and O. Mutlu. Evaluating STT-RAM as an energy-efficient main memory alternative. In *IEEE Int. Symp. on Performance Analysis of Systems and Software (ISPASS)*, pages 256–267, April 2013.

[KSS12] Anil Krishna, Ahmad Samih, and Yan Solihin. Data sharing in multi-threaded applications and its impact on chip design. In *Int. Symp. Performance Analysis of Systems and Software (ISPASS)*, pages 125–134. IEEE, 2012.

[LCG+16] David Lo, Liqun Cheng, Rama Govindaraju, Parthasarathy Ranganathan, and Christos Kozyrakis. Improving resource efficiency at scale with Heracles. *ACM Transactions on Computer Systems (TOCS)*, 34(2):6, 2016.

[LK14] Jacob Leverich and Christos Kozyrakis. Reconciling high server utilization and sub-millisecond

quality-of-service. In *Proceedings of the Ninth European Conference on Computer Systems*, page 4. ACM, 2014.

[LTCS10] M. A. Laurenzano, M. M. Tikir, L. Carrington, and A. Snavely. PEBIL: Efficient static binary instrumentation for Linux. In *IEEE Int. Symp. on Performance Analysis of Systems Software (ISPASS)*, pages 175–183, March 2010.

[MHSN15] D. Molka, D. Hackenberg, R. Schone, and W. E. Nagel. Cache Coherence Protocol and Memory Performance of the Intel Haswell-EP Architecture. In *Int. Conf. on Parallel Processing (ICPP)*, pages 739–748, Sept 2015.

[MSM+11] Sai Prashanth Muralidhara, Lavanya Subramanian, Onur Mutlu, Mahmut Kandemir, and Thomas Moscibroda. Reducing memory interference in multicore systems via application-aware memory channel partitioning. In *Proc. 44th IEEE/ACM Int. Sym. Microarchitecture*, MICRO-44, pages 374–385. ACM, 2011.

[PB14] A. J. Pena and P. Balaji. Toward the efficient use of multiple explicitly managed memory subsystems. In *IEEE Int. Conf. on Cluster Computing (CLUSTER)*, pages 123–131, Sept 2014.

[QP06] Moinuddin K. Qureshi and Yale N. Patt. Utility-based cache partitioning: A low-overhead, high-performance, runtime mechanism to partition shared caches. In *Proc. 39th IEEE/ACM Int. Symp. Microarchitecture*, MICRO 39, pages 423–432. IEEE Computer Society, 2006.

[RKB+09] Brian M Rogers, Anil Krishna, Gordon B Bell, Ken Vu, Xiaowei Jiang, and Yan Solihin. Scaling the bandwidth wall: challenges in and avenues for CMP scaling. *ACM SIGARCH Computer Architecture News*, 37(3):371–382, 2009.

[SHF+15] Christopher Sewell, Katrin Heitmann, Hal Finkel, George Zagaris, Suzanne T Parete-Koon, Patricia K Fasel, Adrian Pope, Nicholas Frontiere, Li-ta Lo, Bronson Messer, et al. Large-scale compute-intensive analysis via a combined in-situ and co-scheduling workflow approach. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC'15*, page 50. ACM, 2015.

[TJS09] Kai Tian, Yunlian Jiang, and Xipeng Shen. A study on optimally co-scheduling jobs of different lengths on chip multiprocessors. In *Proc. 6th ACM Conf. Computing Frontiers*, CF '09, pages 41–50. ACM, 2009.

[ZBF10] Sergey Zhuravlev, Sergey Blagodurov, and Alexandra Fedorova. Addressing shared resource contention in multicore processors via scheduling. *ACM Sigplan Notices*, 45(3):129–142, 2010.

[ZHG+15] H. Zhu, L. He, B. Gao, K. Li, J. Sun, H. Chen, and K. Li. Modelling and developing co-scheduling strategies on multicore processors. In *44th Int. Conf. Parallel Processing (ICPP)*, pages 220–229. IEEE Computer Society, 2015.

[ZLMT14] Yunqi Zhang, Michael A Laurenzano, Jason Mars, and Lingjia Tang. Smite: Precise QOS prediction on real-system SMT processors to improve utilization in warehouse scale computers. In *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 406–418, 2014.

Guillaume Aupy is a tenured researcher at Inria Bordeaux Rhônes-Alpes. He received his PhD from École Normale Supérieure de Lyon (ENS Lyon) in 2014 where he worked on reliable and energy efficient scheduling for High-Perfomance Computers (HPC). He was a research assistant professor at Penn State University in 2015 then in Vanderbilt University in 2016 where he worked on the impact of data-management in HPC. His recent research interests include IO management in HPC systems. He is also interested in scheduling techniques and parallel algorithms for distributed systems, energy-aware and fault-tolerant algorithms. He is the technical program vice-chair for SC'2017 and area chair for ICA3PP'17.

Anne Benoit received the PhD degree from Institut National Polytechnique de Grenoble in 2003, and the Habilitation à Diriger des Recherches (HDR) from École Normale Supérieure de Lyon (ENS Lyon) in 2009. She is currently an associate professor in the Computer Science Laboratory LIP at ENS Lyon, France. She is the author of 38 papers published in international journals, and 78 papers published in international conferences. She is the advisor of 8 PhD theses. Her research interests include algorithm design and scheduling techniques for parallel and distributed platforms, and also the performance evaluation of parallel systems and applications, with a focus on energy awareness and resilience. She is Associate Editor of IEEE TPDS, JPDC, and SUSCOM. She is the program chair of several workshops and conferences, in particular she is program chair for HiPC'2016, program co-chair for

ICPP'2017, and technical papers chair for SC'2017. She is a senior member of the IEEE, and she has been elected a Junior Member of Institut Universitaire de France in 2009.

Sicheng Dai received his BS in 2015, from School of Information Science and Engineering, Lanzhou University, Gansu, China. He now is a MS student in Computer Science and Software Engineering, East China Normal University. As part of his studies, he spent three months as visiting student at ENS Lyon, working on co-scheduling.

Loïc Pottier completed his master at the University of Versailles in 2015, and then moved to École Normale Supérieure de Lyon (ENS Lyon), where he is currently a PhD candidate under the supervision of Anne Benoit and Yves Robert. As part of completing his PhD, he also spent three months as visiting student at Argonne National Laboratory, where he worked with Swann Perarneau. His main topics of interest include co-scheduling, fault tolerance, and scheduling techniques for large scale platforms.

Padma Raghavan is the Vice Provost for Research and Professor of Computer Science and Computer Engineering at Vanderbilt University. Prior to joining Vanderbilt in 2016, she served as the Associate Vice President for Research and Strategic Initiatives, as the founding Director of the Institute for CyberScience and Distinguished Professor of Computer Science and Engineering at the Pennsylvania State University. Raghavan specializes in high-performance computing and its applications with a particular focus on sparse graph and matrix problems. She has supervised many M.S and Ph.D, theses and authored over one hundred peer-reviewed publications in three areas: scalable parallel computing; energy-aware supercomputing; and computational modeling, simulation and knowledge extraction. Raghavan currently serves on the Council of SIAM (Society of Industrial and Applied Mathematics), its committee on Science Policy and the editorial boards of its series on Computational Science and Engineering, the SIAM Series on Software, Environments and Tools, She co-chairs the Technical Program at Supercomputing 2017. Raghavan is a Fellow of the IEEE (Institute of Electrical and Electronics Engineers) and she received the National Science Foundation's CAREER award and the Maria Goeppert-Mayer Distinguished Scholar award from the University of Chicago and the Argonne National Laboratory, in recognition of her contributions to scalable parallel computing.

Yves Robert received the PhD degree from Institut National Polytechnique de Grenoble. He is currently a full professor in the Computer Science Laboratory LIP at ENS Lyon. He is the author of 7 books, 147 papers published in international journals, and 219 papers published in international conferences. He is the editor of 11 book proceedings and 13 journal special issues. He is the advisor of 30 PhD theses. His main research interests are scheduling techniques and resilient algorithms for large-scale platforms. Yves Robert served on many editorial boards, including IEEE TPDS and JPDC. He was the program chair of HiPC'2006 in Bangalore, IPDPS'2008 in Miami, ISPDC'2009 in Lisbon, ICPP'2013 in Lyon and HiPC'2013 in Bangalore. He is a Fellow of the IEEE. He has been elected a Senior Member of Institut Universitaire de France in 2007 and renewed in 2012. He has been awarded the 2014 IEEE TCSC Award for Excellence in Scalable Computing, and the 2016 IEEE TCPP Outstanding Service Award. He holds a Visiting Scientist position at the University of Tennessee Knoxville since 2011.

Manu Shantharam is a Computational Scientist in the San Diego Supercomputer Center. Manu received his Ph.D. in Computer Science and Engineering from The Pennsylvania State University in 2012. His research interests include sparse scientific computations, scheduling HPC workloads, and resiliency and performance analysis of HPC applications.